THE EPIC 400 ANALYSIS

Jing Hua Zhao
MRC Epidemiology Unit
Strangeways Research Laboratory
Worts Causeway
Cambridge CB1 8RN
Comments sent to jinghua.zhao@mrc-epid.cam.ac.uk

Cambridge 10/5/2006

MRC Epidemiology Unit

Introduction

- The data were from a whole-genome association (WGA) study of breast cancer, from the Genetic Epidemiology Unit.
- The initial purpose of the project was to provide an exercise for our WGA, the sample involve 400 controls (all females) and >220k SNPs, the main phenotype of interest is BMI.
- Preliminary results were reported by Ruth on the Unit Away Day (May 3); this is somewhat a technical but hopefully extended discussion.

Why such an exercise necessary?

- The WGA poses statistical and computational challenges (Carlson et al. (2004) Nature; Hirschhorn & Daly (2005), Wang et al. (2005) Nat Rev Genet)
- The requirement of resources (space and time) due to large number of SNPs is phenomenal.
- Other statistical issues such as multiple testing, complex models of haplotypes have only been widely aware recently.
- The study therefore is a concrete example for our WGA.

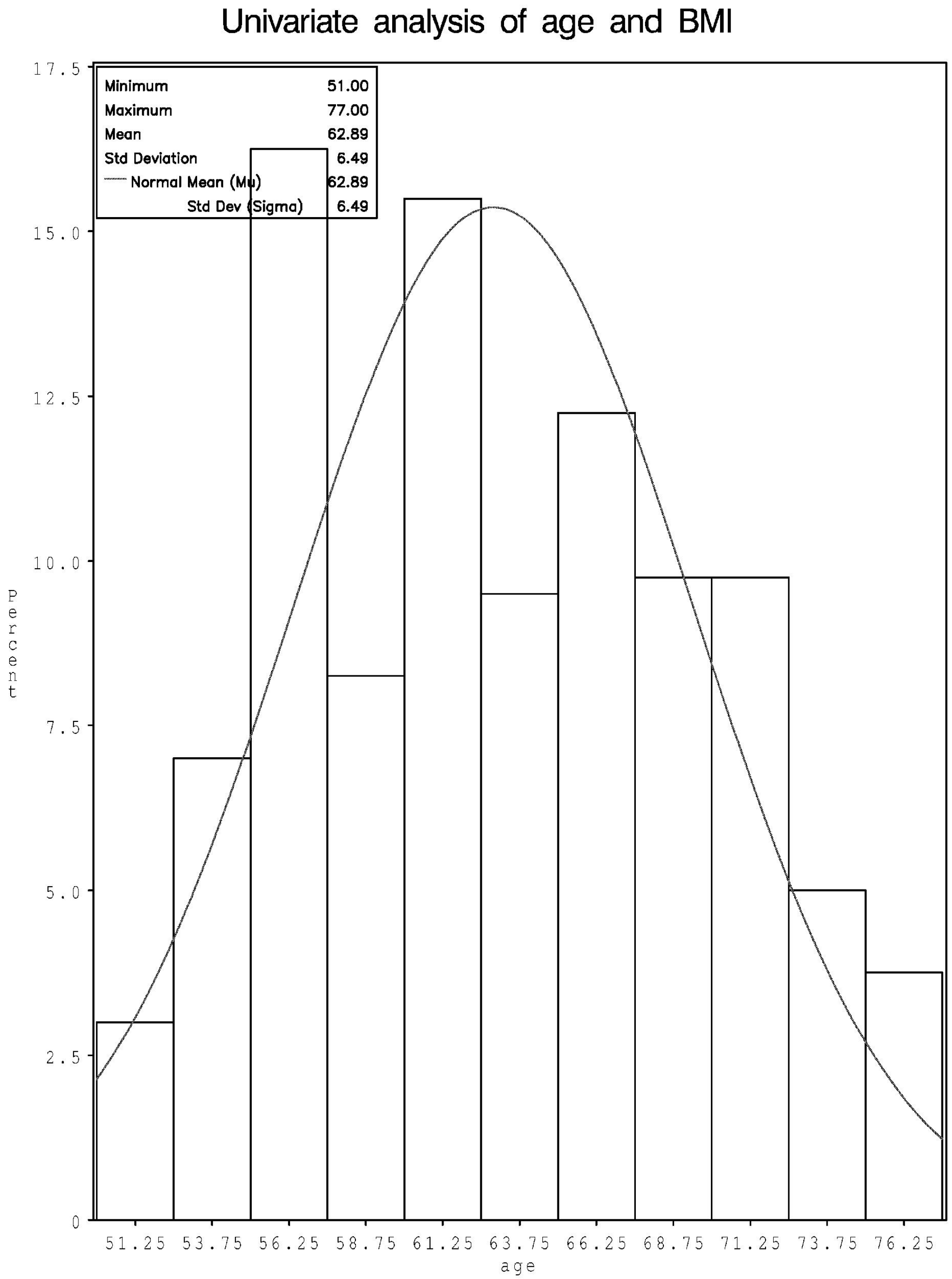
The strategy for data analysis

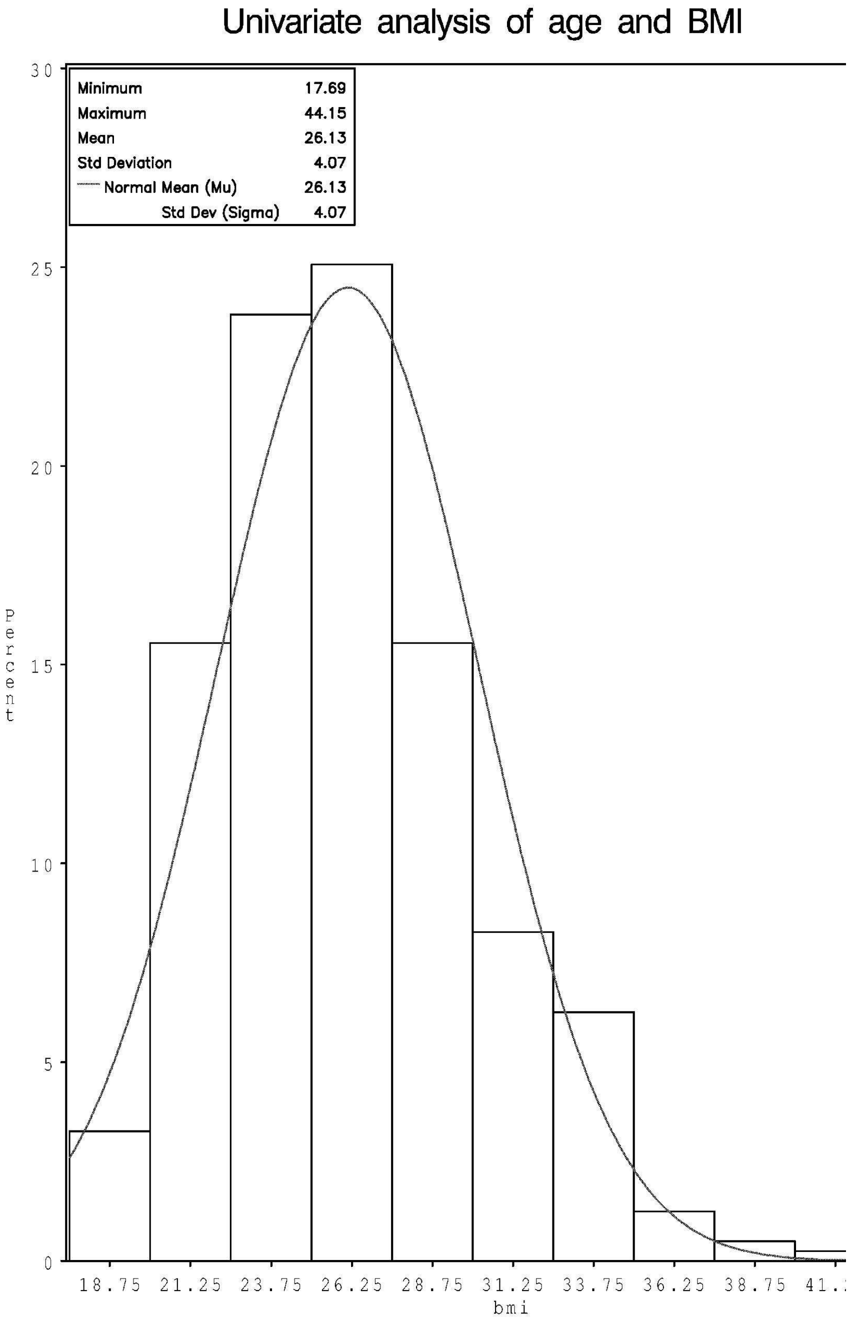
- Information involved (map, genotype, phenotype) is stored in a long, skinny format.
- How to use these information?
  - There is a large number of computer programs developed each with own scope, problems and limitations (Salem et al. (2005) Hum Genomics; Zhao & Tan (2006) Curr Bioinformatics) but less tested (Spence et al. 2003).
  - SAS/GENETICS is used, chosen in view of the limitations reported; especially all results can be directed to datasets with ODS.
  - The analyses include call rates, HWE, single-point analysis.

The set of SAS programs

- **v:\jinghua\EPIC400\readme.txt** descibes the following SAS programs: map.sas, genotype.sas, epic.sas, call.sas, hwe.sas, regress.sas, multtest.sas, pplot.sas, haploview.sas.
- Same code can be used in batch mode under Linux and Windows, by different users and on different phenotypes (e.g. through SAS macro)
- sas.bat has also been used in Soren's problem, saving log/print overflow

MRC Epidemiology Unit


Univariate analysis of age and BMI
Minimum 51.00
Maximum 77.00
Mean 62.89
Std Deviation 6.49
Normal Mean (Mu) 62.89
Std Dev (Sigma) 6.49
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
Percent
51.25
53.75
56.25
58.75
61.25
63.75
66.25
68.75
71.25
73.75
76.25
age


Univariate analysis of age and BMI
Minimum 17.69
Maximum 44.15
Mean 26.13
Std Deviation 4.07
Normal Mean (Mu) 26.13
Std Dev (Sigma) 4.07
30
25
20
15
10
5
0
Percent
18.75
21.25
23.75
26.25
28.75
31.25
33.75
36.25
38.75
41.
bmi

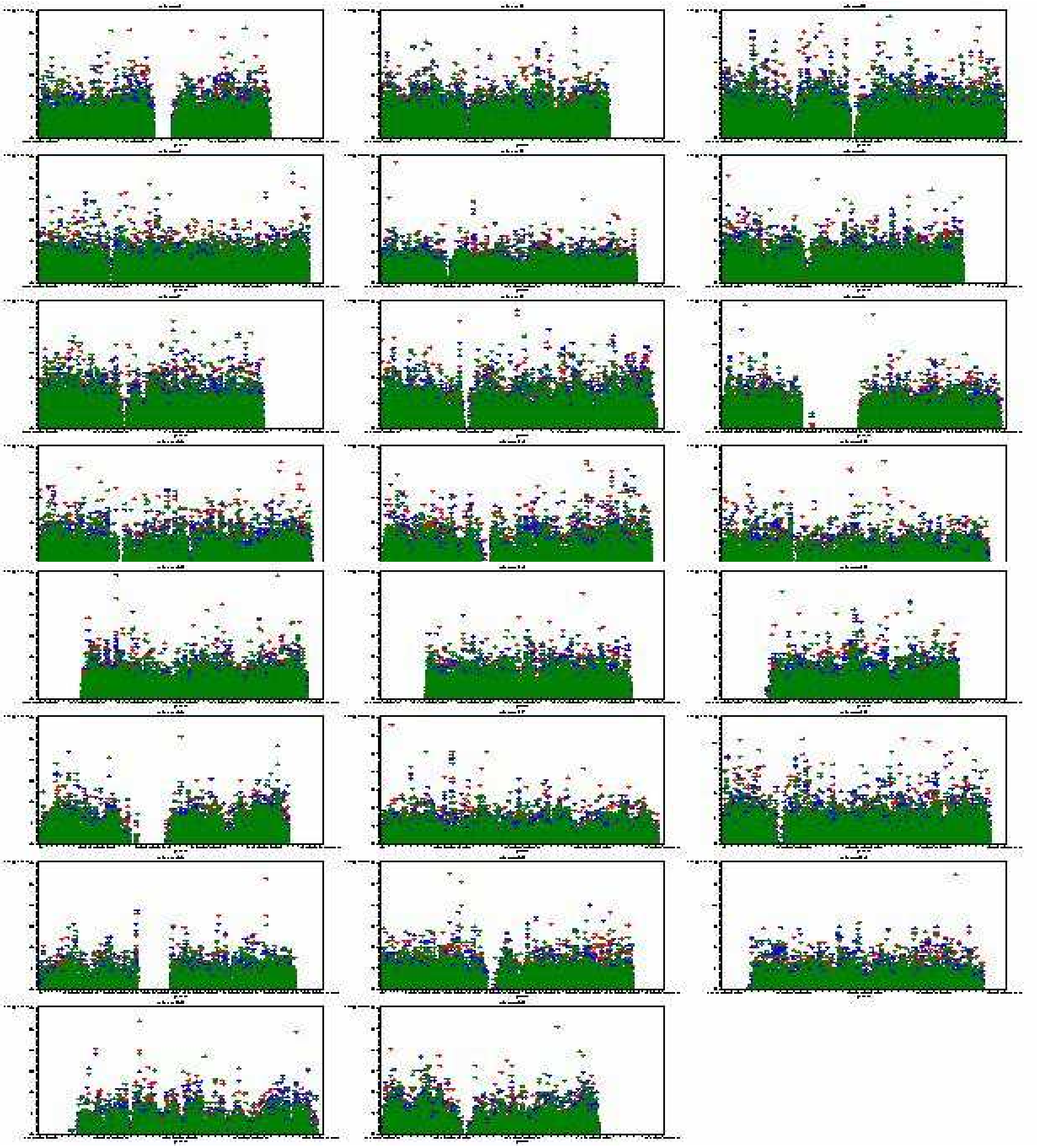
Distribution of SNPs by chromosome and HWD

Chromosome	No	Yes	Total	Acc. frequency	Percent
1	15511	1735	17246	17246	0.08
2	16780	1836	18616	35862	0.16
3	13912	1415	15327	51189	0.22
4	12956	1400	14356	65545	0.29
5	13440	1436	14876	80421	0.35
6	12712	1396	14108	94529	0.41
7	11008	1227	12235	106764	0.47
8	11225	1279	12504	119268	0.52
9	9124	976	10100	129368	0.57
10	10449	1115	11564	140932	0.62
11	9625	1033	10658	151590	0.67
12	9625	1045	10670	162260	0.71
13	7979	889	8868	171128	0.75
14	7078	769	7847	178975	0.79
15	6513	678	7191	186166	0.82
16	6443	667	7110	193276	0.85
17	5493	600	6093	199369	0.87
18	6517	681	7198	206567	0.91
19	3312	381	3693	210260	0.92
20	5069	598	5667	215927	0.95
21	3140	321	3461	219388	0.96
22	2837	343	3180	222568	0.98
23	4801	485	5286	227854	1.00
Total	205549	22305	227854		

Note. The Hardy-Weinberg disequilibria may also be caused by monomorphic SNPs, not entirely due to genotype errors.

Note. In the following slide, genotype is defined as 0=l/l, 1=l/u, 2=u/u, dominant as 0=l/l, 1=(l/u, u/u) and recessive as 0=(l/l, l/u), 1=u/u. SNP rs7721337 at chromosome 5 appears to be outstanding with FDR = 0.015. This is in line with report by Herbert et al. (2006) Nat Genet!

chr	pos	rsn	l	u	Model	Estimate	StdErr	t Value	Std Est	Probt
1	76311117	rs289698	A	C	rec	18.1	4.0	4.54	0.22197	7.61E-06
1	96176533	rs4950091	A	G	rec	2.7	0.6	4.69	0.22924	3.72E-06
1	1.62E+08	rs16840116	C	T	rec	18.1	4.0	4.54	0.22197	7.61E-06
1	1.96E+08	rs1368986	C	T	rec	4.3	0.9	4.62	0.22589	5.19E-06
1	2.19E+08	rs3008633	C	T	rec	9.3	2.0	4.63	0.22651	4.88E-06
1	2.42E+08	rs1715780	C	T	rec	3.6	0.8	4.5	0.2205	8.77E-06
2	2.06E+08	rs1540364	C	T	gid	2.3	0.5	4.86	0.2372	1.65E-06
2	2.06E+08	rs1540364	C	T	dom	2.5	0.5	4.73	0.23113	3.08E-06
3	1.18E+08	rs4855915	C	G	dom	1.9	0.4	4.62	0.22584	5.21E-06
4	1.81E+08	rs7670329	A	G	gid	1.9	0.4	4.73	0.23081	3.18E-06
4	1.81E+08	rs7670329	A	G	dom	2.2	0.5	4.51	0.22059	8.69E-06
5	4721289	rs4407602	A	G	rec	4.0	0.9	4.72	0.23051	3.28E-06
5	10283714	rs7721337	A	T	rec	13.0	2.3	5.7	0.27507	2.33E-08
5	65703234	rs982965	A	T	dom	1.9	0.4	4.53	0.22147	7.98E-06
5	1.44E+08	rs17096446	A	G	rec	18.1	4.0	4.54	0.22197	7.61E-06
6	3252025	rs17246629	C	G	rec	18.1	4.0	4.54	0.22197	7.61E-06
6	66748379	rs16897624	A	G	rec	18.1	4.0	4.54	0.22197	7.61E-06
9	8858177	rs2217711	C	T	gid	1.7	0.4	4.62	0.22603	5.12E-06
9	10629472	rs16926200	A	G	rec	13.8	2.8	4.91	0.23939	1.32E-06
9	74672978	rs17060568	A	G	rec	9.5	2.0	4.78	0.23316	2.5E-06
12	63264342	rs17223200	A	G	rec	18.1	4.0	4.54	0.22197	7.61E-06
12	80213806	rs7958510	A	G	rec	13.1	2.8	4.65	0.22731	4.51E-06
13	32650866	rs797215	C	T	gid	-1.4	0.3	-4.53	-0.22152	7.95E-06
13	1.02E+08	rs16960116	C	T	rec	14.0	2.8	5	0.24342	8.6E-07
14	85817874	rs1394702	G	T	rec	10.4	2.3	4.5	0.22013	9.09E-06
15	24884556	rs17739073	A	G	rec	18.1	4.0	4.54	0.22197	7.61E-06



Plot of $\log_{10}p_t$ by chromosomes

Red=Dominant

Blue=Additive

Green=Recessive

What does the exercise tell us?

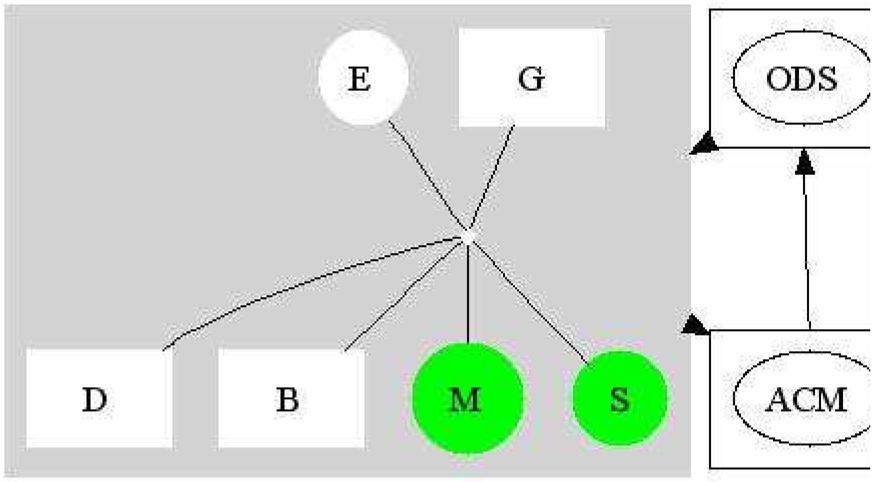
- There are 436 individuals at phase 1 and 400 individuals with genotype data. The number of SNPs is 227,874. The gzipped ASCII file was about 350MB or 4GB, while the final SAS dataset was about 10GB. The working directory is about 20GB. The processing time varies for different tasks (call rates, Hardy-Weinberg tests, regression, adjustment for multiple testing, etc) but on average finishes within half a day. There were problems with the initial setup for SAS, but it appeared to be ok after further space was allocated for swapping and sorting. Note that Hardy-Weinberg tests would not work under Windows (2GB RAM), possibly needs sole use of the Windows system.
- It is therefore expected that for the WGA study of obesity, the size of problem would be approximately ~30 times larger at the first stage (~5,000 individuals, ~600,000 SNPs).
- What is the best infrastructure? Examples are www.genesniffer.org, Genomizer, GMED, information as from Integragen, etc.

MRC Epidemiology Unit

Other aspects

- It took over six months to obtain EPIC 400 data; their 2nd stage data is not available to us.
- We should obtain phenotype data of EPIC 500 now to check for “incident” cases. The EPIC 5000 is not an exact replica of EPIC 400; cohort methods are required rather than the “standard” case-control paradigm.
- InterAct is another general, sophisticated framework, coupling with non-genetic factors.
- Additional requirement would include covariate information and data from associate projects such as the 1958 cohort, HapMap and Genetic Analysis Workshops (GAWs). These would allow for a broader range of analyses be conducted.
- The major benefit of SAS/GENETICS also lies on the ease to share work. However, it is relatively recent so alternatives are still viable for some analyses. The framework would somewhat be applicable to other analyses in the unit in general.

So, the next steps



- To establish an integrated system – including data from many other sources; it will particularly be useful to study software tools and information by our genotypers
- The computational platform is preferably a Linux cluster, such that each node can run one chromosome. If the tasks can be allocated to a 30-node Linux cluster, the CPU time is expected to be similar. More sophisticated modelling is possible. Similar architectures have been used in the Wellcome Trust Centre for Human Genetics, Oxford, and elsewhere for years. Another recent example on Bayesian admixture mapping was reported in the 4th UK Diabetes Genetics Consortium Meeting, Cambridge (P McKeigue).
- Some detailed analysis can be carried out in other software systems such as S-Plus/R/Stata (e.g. Stata/MP) – the current HTR is inappropriate for GxE interaction but haplo.stats in S-Plus/R.
- I believe this is advantageous than many other approaches that sketched in the literature.

MRC | Epidemiology Unit

Software systems (Windows/Linux)

- **SAS** is a well-established and powerful package system
- **Stata**
  - Currently, phase inference is often through other programs such as SNPHAP with the posterior probabilities as sampling weight in the analysis
- **S-PLUS/R**
  - gap, haplo.score, haplo.stats, hapassoc, haplo.cc
  - haplotype clustering of Tzeng et al. (2006) Am J Hum Genet

SAS

- SAS/GENETICS was mainly designed for association tests, including procedures ALLELE, CASE-CONTROL, HAPLOTYPE, HTSNP, FAMILY, PSMOOTH, MULTTEST, INBREED
- Powerful database management facility and graphics
- The variety of procedures is particularly attractive if covariates are involved, e.g. **haplotype trend regression**, but less inclusive (e.g. X-chromosome data)

Stata

- A recent review by Hilbe (2006) Amer Stat
- It facilitates extension, easy installation and on-line documentation, examples include,
  - CIMR site (http://www-gene.cimr.cam.ac.uk)
  - Biostatistics Resources (http://www.biostat-resources.com)
  - Others listed at http://slack.ser.man.ac.uk
- It has unique features appropriate for haplotype analysis, so far slightly cumbersome to prepare data
- It is possible to develop programs through plugins

S-PLUS/R

- S-PLUS
  - Known for its object-oriented programming language and powerful graphics
  - Limited number of packages including haplo.score, haplo.stats, multigene, kinship
- R
  - an integrated environment for statistical computing and genetic data analysis
  - compatible with S-PLUS but with more variety of packages, now over 700
  - Freely available

Tools for haplotype-traits association

- Zhao et al. (2000) Hum Hered
  - EHPLUS/GENECOUNTING, global+haplotype specific tests
- Zaykin et al. (2002) Hum Hered
  - HTR, Haplotype trend regression
- Schaid et al. (2002), Lake et al. (2003), Burkett et al. (2004) Hum Hered
  - haplo.score, haplo.stats and hapassoc, generalised linear model (GLM) framework
- Zhao et al. (2003) Am J Hum Genet
  - Hplus, generalised estimation equations (GEE) framework
- Epstein & Satten (2003) Am J Hum Genet
  - CHAPLIN, logistic regression
- Tzeng et al. (2006) Am J Hum Genet
  - R program, GLM with haplotype clustering

An example of PROC HAPLOTYPE

```
proc haplotype data=ccc out=outhap;
        var m1-m12;
run;
proc print data=outhap;
    title ‘Marker-marker analysis with haplotype assignments’;
run;
proc haplotype data=ccc noprint;
    var m1-m12;
    trait cc / testall perms=1000;
run;
proc print data=outhap noobs round;
    title’Tests for haplotype-trait association’;
run;
```

Tests for haplotype-trait association

Global tests for haplotype-trait association

```
 Trait    Trait           Num                                 Chi-     Pr >       Prob
Number    Value           Obs        DF       LogLike       Square     ChiSq     Exact

          1            1       871          49         -1779
          2            0      1257          49         -2479
               Combined       2128          49         -4269      20.2716    0.9999    0.3740
```

Haplotype-specific tests

```
                             ----------Frequencies---------
                                                                    Chi-      Pr >      Prob
    Number    Haplotype       Trait1     Trait2    Combined       Square     ChiSq     Exact

         1    1-1-1-1-1-1    0.58563    0.62571     0.60931       6.9463    0.0084    0.0080
         2    1-1-1-1-1-2    0.00000    0.00000     0.00000            0    1.0000    1.0000
```

The results include an omnibus heterogeneity test, followed by simple proportion tests for individual haplotypes as would be from EHPLUS and GENECOUNTING (if missing data is used)

Advantages and disadvantages

- Global test serves as a universal test base but only applies to categorical outcome and do not account for covariates, HWE
- Score statistics can be rapidly calculated along with simulation but not allow for nested models
- HTR is applicable to large number of designs
- Current retrospective methods do not account for covariates
- The GLM and GEE frameworks allow for G x E interactions but need more assumption (GEE) or more involved

Power considerations

- Prospective (Schaid et al. (2002) Am J Hum Genet) versus retrospective likelihood (Epstein & Satten (2003) Am J Hum Genet; Tan et al. (2005) Hum Hered; Tan et al.(2005) Genet Res)
- Simulation
  - Satten & Epstein (2004) Genet Epidemiol:
    - Comparable performance for multiplicative model
    - Retrospective likelihood more powerful for dominant/recessive models
- Empirical studies
  - de Bakker et al. (2005) Nat Genet
  - van Steen et al. (2005) Nat Genet
- Multiple imputation is applicable to all methods

Statistical significance

- Potentially, there is a large number of degrees of freedom
- Monte Carlo method or permutation test is often necessary
- Adjustment for p-value is often required in view of the multiple-testing involved

Multiple testing

- Methods
  - Bonferroni
  - Sidak
  - Hockberg
  - Holm
  - FDR
- Software systems
  - SAS/STAT PROC MULTTEST
  - Stata package smile (Roger et al. (2003) Stata J, st0035)
  - R packages (multtest, qvalue, etc)

G x E interactions

- The null hypothesis of no association
  - Zaykin et al. (2002), Schaid et al. (2002) Hum Hered
- The alternative hypothesis
  - Lake et al. (2003) Hum Hered
  - Zhao et al. (2003) Am J Hum Genet
- Confusion in the literature: haplotype estimates are based on the null hypothesis
  - Dong et al. (2004) Hypertension

Examples of ODBC/MySQL in SAS

```
libname test odbc datasrc=myodbc user=jhz22;
proc print data=test.Genotypes_chr11_CEU;
run;
proc sql;
        connect to odbc as test (datasrc=myodbc user=jhz22);
        select * from test.pet;
libname test2 mysql database=test user=jhz22;
proc print data=test2.pet;
run;
proc sql;
        connect to mysql as test2 (database=test user=jhz22);
        select * from test2.pet;
```

An example of ODBC in Stata

```
. odbc list
. odbc query "MySQL database“
. odbc desc "Genotypes_chr4_HCB", dialog(complete)
. set mem 50M
. odbc load, exec("select * from Genotypes_chr4_HCB")
```

MRC Epidemiology Unit

An example of RODBC

```
# load ODBC
library(RODBC)
c2 <- odbcConnectAccess(“db1.mdb”)
# select the table
tblOutput <- sqlQuery(c2,paste(“select * from Genotypes_chr11_CEU”))
# the property of tblOutput
class(tblOutput)
```

q-values

```
library(qvalue)
pvalues <- scan("ant.out")
# qvalues <- qvalue(pvalues)
# qwrite(qvalues,"qvalues.txt")
#  necessary due to the U shape of the p-value histogram
qvalues.boot <- qvalue(pvalues,pi0.method="bootstrap")
# fdr.level=0.05
plot(qvalues.boot)
qsummary(qvalues.boot,cut=c(0.0001,0.001,0.01,0.05))
```

q-values

```
Call:
qvalue(p = pvalues, pi0.method = "bootstrap")

pi0:    0.41

Cumulative number of significant calls:

        <1e-04 <0.001 <0.01 <0.05
p-value     88    231   611  1042
q-value      2     58   407   990
```

pi0 is the true negative rate using observed p values which exceeds a value lambda

Further information

- **SAS** (http://www.sas.com) fully documented and a macro for haplotype trend regression available
- **Stata** (http://www.stata.com)
- **S-PLUS** (http://www.insightful.com)
- **R** (http://cran.r-project.org)
- Linkage server (http://linkage.rockefeller.edu)

Acknowledgements

- Ruth Loos
- Jian’an Luan
- Lisa Purlow
- Iain Morrison
- Jonathan Morrison

MRC Epidemiology Unit